Expectation Maximization

Given dataset $D = \{(x_n)_{n=1}^N\}$ and given

$$P(x) = \sum_{k=1}^{k} \pi_k \, N(x; \mu_k, \Sigma_k)$$

x is a vector in an N-dimensional space
μ_k is a vector in a k-dimensional space
Σ_k is a matrix $\in \mathbb{R}^d$ where d is the dimension of the points x

determine μ_k, Σ_k, π_k → $\forall$ gaussian compute prior, mean and covariance matrix

This algorithm is a generalization of K-means. → only computes μ_k

This is an iterative algorithm that computes the maximum likelihood solution

$$\underset{\pi,\mu,\Sigma}{\text{argmax}} \ln P(x \mid \pi, \mu, \Sigma)$$

likelihood is probability of x given the parameters of the model
maximum likelihood solution are the parameters for which $P(x)$ is maximized

This mathematical formulation has close form solution:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

dependence on the dataset
dependence on the posterior on z

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N} \quad \text{with} \quad N_k = \sum_{n}^{N} \gamma(z_{nk})$$

This means that if you have the dataset and the posterior you can compute the optimal parameters.
We don't have the posterior.

We can assume to have the optimal parameters in order to compute the posterior $\gamma(z_k)$ instead.

posterior $\gamma(z_k) = P(z_k = 1 \mid x) = \frac{P(z_k = 1) P(x \mid z_k = 1)}{P(x)}$ Bayes rule

$$= \frac{\pi_k N(x; \mu_k, \Sigma_k)}{\sum_{j=1}^{N} \pi_j N(x; \mu_j, \Sigma_j)}$$

So, we have to paths:

given π_k, μ_k, Σ_k compute $\gamma(z_k)$ → Expectation step

given $\gamma(z_k)$ compute π_k, μ_k, Σ_k → Maximization step

Expectation Maximization just repeats these two steps in an iterative way. This is done until convergence

Algorithm

① Initialize $\pi_k^{(0)}, \mu_k^{(0)}, \Sigma_k^{(0)}$ → start with a random initialization

② Repeat until termination condition $t = 0, \dots, T$

① E step → given $\pi_k^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)}$ compute $\gamma(z_{nk})^{(t+1)}$

$$\gamma(z_{nk})^{(t+1)} = \frac{\pi_k^{(t)} \mathcal{N}(x_n \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(x_n \mid \mu_j^{(t)}, \Sigma_j^{(t)})}$$

② M step → given $\gamma(z_{nk})^{(t+1)}$ compute $\pi_k^{(t+1)}, \mu_k^{(t+1)}, \Sigma_k^{(t+1)}$

$$\mu_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})^{(t+1)} x_n$$

$$\Sigma_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})^{(t+1)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T$$

$$\pi_k^{(t+1)} = \frac{N_k}{N} \quad \text{with} \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})^{(t+1)}$$

If you use a fixed covariance then the algorithm reduces to k means. The E step in K-means correspond to the step in which you associate each point to the closest centroid and form the cluster.

Robust to outliers when using gaussian distribution but convergence is slower.

Remarks

① EM converges to <u>local</u> maximum likelihood
② EM is an extension of K-means that take into account covariance
③ Can be generalized to other distributions other than Gaussian

Generalized Expectation Maximization Method

We have :

- Observed data $X = \{x_1, \dots, x_N\}$
- Unobserved latent values $Z = \{z_1, \dots, z_N\}$
- θ set of parameters
- $Y = \{(x_n, z_n)_{n=1}^N\}$ the full data

We want to determine :

- θ^* that locally maximizes $E[\ln P(Y|\theta)]$

General EM algorithm

$Q(\theta'|\theta)$ likelihood function : computes θ' from θ

① E step : calculate θ' given θ and X $\quad Q(\theta'|\theta) \longleftarrow E[\ln P(Y|\theta')|\theta, X]$

② M step : replace hypothesis θ by the hypothesis θ' $\quad \theta \longleftarrow \underset{\theta'}{\operatorname{argmax}}\, Q(\theta'|\theta)$